Department of Computer Science and Engineering, IIT Palakkad

CS5512 Machine Learning

EndSem (30 June, 2019)

Time: 11:00am to 12:30 pm Points: 20

Instructions

1. There are a total of **4 questions** in **2 pages** in this Question paper.
2. Write your **name** and **roll number** on the answer sheet.
3. Answers for all parts of a question should be together. If different parts of a question is scattered in answer sheet then only the first part of it will be evaluated.
4. Write your answers neatly in blue or black **ink**. Do not use pencil or red ink. If the answer is not legible, benefit of doubt will go to the wrong answer.

1. (a) Write two main differences between clustering outputs of K-means and Gaussian Mixture Model. [2]

   (b) Suppose you are free to chose any distance measure other than Euclidean distance for K-means Algorithm.

   (a) Name a distance measure that you will choose to get ellipsoidal shaped clusters from K-means? Justify your answer ? [1]

   (b) Write the algorithm of your modified kmeans in details with formulation for estimation of all kind of parameters of that distance function, cluster centroids, clusters assignment. [4]

2. Consider the spectral clustering method to answer the following

   (a) What is a graph cut? [hint: Give a mathematical definition of graph cut for two partitions.] [1]

   (b) What is Normalized Cut? [hint: Give a mathematical definition of normalized graph cut for two partitions.] [1]

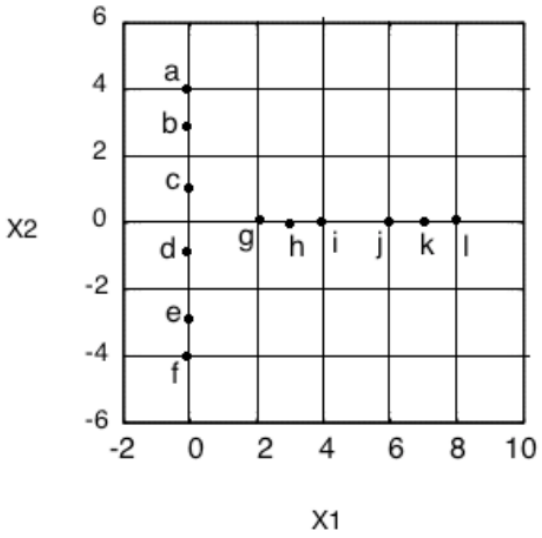
   (c) What is the significance of normalized cut with respect to the spectral clustering? [1]

   (d) Consider the two partitioning cases and prove or disprove that finding minimum cut is equivalent to solve the following [3]

   $$\min_f f^T L f,$$

   where L denotes the graph Laplacian and f is suitable partition index as discussed in class. Please define all notation you use in your proof including f and L.

3. Consider the data set given in the following figure and various clustering methods. Give possible clusters as output for



(a) K-means clustering with $K = 2$. [1]

(b) DBSCAN clustering with radius = 0.51 and MinPts=2 (including the point itself). [2]

(c) One can produce same clusters using hierarchical clustering (agglomerative) on the above data set. Draw the cluster tree (a dendrogram) for following linkages and write if clusters (considering number of clusters =2) from dendrogram are similar to the clusters achieved of previous two methods

(a) Single Linkage [2]

(b) Complete Linkage [2]

4. Prove that the VC dimension of a linear classifier in $\mathbb{R}^d$ space is equal to $d + 1$. [**This is a bonus question and has no partial marks. This question has 5 marks**]